Réduction de Données pour une agriculture intelligente

Christian Salim, Nathalie Mitton

Réduction de Données pour une agriculture intelligente

Christian Salim et Nathalie Mitton

Inria, France

De nos jours, le domaine de l'agriculture est confronté à de nombreux défis pour une meilleure utilisation de ses ressources naturelles. Pour cette raison, il est nécessaire de localement superviser les données météorologiques et les conditions du sol pour prendre des décisions mieux adaptées à chaque culture. Les réseaux de capteurs sans fil (RCSF) peuvent servir comme système de surveillance pour ces types de paramètres mais les nœuds capteurs ont des ressources matérielles et énergétiques limitées. Le processus d'envoi d'une grande quantité de données au puits entraîne une grande consommation d'énergie et une utilisation importante de la bande passante. Dans ce papier, nous proposons un algorithme de réduction de données utilisant le coefficient de corrélation de Pearson (PDCP) pour prédire les nouvelles valeurs au niveau du nœud capteur et du puits. Cette approche est validée par des simulations sur MATLAB tout en utilisant des ensembles de données ouvertes Weather Underground. Les résultats valident l'efficacité de notre approche montrant une réduction de données allant jusqu'à 69% tout en maintenant la précision des informations. La prédiction des valeurs d'humidité à partir de la température présente un écart par rapport à la valeur réelle inférieur à 7%.

Mots-clefs : Corrélation, Pearson, Réduction de Données, Prédiction, Smart Agriculture, Réseau de capteurs sans fil

1 Introduction

Modern agricultural fields are in need of new and improved methods to deal with climate change and scarcity of water. In our scenario, we assume a WSN deployed for smart agriculture, periodically gathering environmental data from different sensor nodes and sending this data to a sink for further analysis [ea15] at regular pace. This periodic cycle leads to a lot of redundant data sent to the sink, especially if no changes occur in the monitored feature (e.g. if the temperature stays stable). To reduce the amount of data transmitted by the sensor nodes and thus their energy consumption, in this paper, we introduce a Pearson Data Correlation and Prediction algorithm (PDCP), a data reduction technique based on a machine learning process to predict the data correlation between the same parameter on different neighbour nodes and between different parameters on a single node.If a high correlation is detected, the nodes send less data to the sink. In the inter-nodes correlation, one of the nodes sends the correlated data, reducing interference at the same time. However a critical threshold between the estimated data and the real one is always present on the sensor node level to detect any absurd change in the values as explained in Section 3.

Matlab simulations show the validity of our approach, by reducing the amount of sent data to the sink outperforming other approaches. For the inter-nodes correlation technique the temperature prediction is accurate and a maximum difference of 1 degree Celsius exists if compared to the real value. The data reduction is huge and reaches 100% for the two days of prediction. For the intra-node correlation, while predicting the humidity from the temperature parameter, we needed to send only 6 values out of the 96 existing sensed values and having a difference less than 7% for the humidity parameter prediction while comparing it to the real sensed value.

2 Background and Related Work

Different data reduction techniques for WSN are present in the literature. In this section we survey some of these approaches while focusing on the data correlation and machine learning techniques.

Data correlation tests the correlation between several characteristics. In [ea17], the authors proposed a correlation system based on a Bayesian inference approach in order to avoid transmitting data that can be reconstructed from other data. Machine learning for data prediction is widely used for data reduction [SM20]. In the dual prediction model, both the sensor node and the sink predict the next values of the monitored feature simultaneously as in [ea18] which uses a machine learning technique and send all the data in the learning phase to the sink. In this approach, the authors detect a trend directly after a single change, which can cause some problems for the learning process and send more values. A lot of approaches were interested in data correlation for this purpose, mainly using the Pearson correlation technique and its derivatives [RC19], the Auto Regression model and the convolutional long short-term memory (LSTM) networks techniques. Convolutional LSTM network is too complex to be embedded on sensors. In this paper, our solution for data reduction at the sensor node is to implement a light data reduction algorithm based on data correlation at the sensor node level to reduce the amount of sent data to the sink.

3 Data Correlation

Our contribution consists of using the Pearson correlation method to reduce the amount of redundant data sent from the nodes to the sink while maintaining the needed accuracy of the sensed data. We will run two data correlation based mechanisms : 1) Inter-nodes data and 2) Intra-node data correlation.

3.1 Inter-Nodes Data Correlation

We use the geometric interpretation of the Pearson correlation. This technique needs several values to detect the percentage of correlation between two parameters of the same type on two different sensor nodes. The correlation coefficient between two vectors X_1 and X_2 is computed as follows :

$$\rho_{X_1X_2} = \cos\theta = \frac{\overrightarrow{X_1}.\overrightarrow{X_2}}{\|X_1\|.\|X_2\|} \times s \tag{1}$$

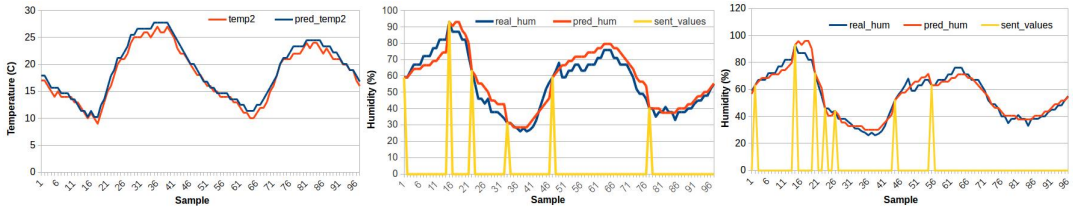
The percentage of correlation is then $\rho_p = \rho^2_{X_1X_2} \times 100$. This method is applied on every sensor node to compute the correlation coefficient with all its neighbour nodes in its communication range. We assume that the information needed to compute the correlation coefficient together with the remaining energy of each node is piggybacked in the Hello messages used for each node to discover each other. A threshold of correlation is predefined on each node for each type of data.The mean ratio of difference R_m is computed between the two parameters after n consecutive values as $R_m = \frac{\sum_{i=0}^{n} \frac{y_i}{x_i}}{n}$. Then, the sink computes the "missing" value of y_{n+1} as follows :$y_{n+1} = y_n + (x_{n+1} - x_n) \times R_m \times \rho^2 \times s$ Once two nodes detect a high correlation between one or several parameters, they locally decide whether to send the message. Only the one with the highest energy sends the message. In case of ties, the message will be sent by the node with the smallest identifier.

3.2 Intra-Node Data Correlation

Data correlation can be computed with different types of data on the same node. For example the correlation between humidity and temperature can be high and follows a certain shape, in this case we can extract one value from the other one. If any important correlation is found, it will help reduce the amount of data sent from the node to the sink by sending only one of the two correlated parameters.

The Pearson Correlation Coefficient is also applied to detect the correlation between two different parameters and start the prediction. The correlation coefficient ρ_{xy} between two different parameters is compared to a predefined threshold of correlation th_{cor}. Two parameters are considered correlated if $\rho_{xy} > th_{cor}$ where x and y are two vectors of several values representing two different parameters. We compute the Pearson correlation coefficient as mentioned in inter-node correlation. If the latter is sufficient, R_m and the next value y_{n+1} are computed from ρ, x_{n+1}, x_n, y_n and R_m.

Meanwhile the sensor nodes keep sensing the real values. The predicted value R_y must be in a certain range based on the real value to be accepted : $th_{low} = (\frac{1-\rho_{xy}}{2} + \rho_{xy}) \times y_r < v < th_{up} = (\frac{1-\rho_{xy}}{2} + 1) \times y_r$. If the predicted value falls outside this range, the node sends the real value to the sink and the prediction process continues based on the new real value.



(a) Temperature at sink (b) Intra node humidity (c) PDCP Humidity prediction

4 Scenario and Algorithm

The PDCP algorithm is detailed in 1. The intra-node correlation is always applicable on each sensor-node, however, the inter-nodes correlation depends on the neighboring parameters (distance and radius of communication). Later on, in our experiments, a full day of values is needed to compute the correlation (48 values are captured in a day by a sensor node for every parameter). In this scenario and algorithm, for the intra-node correlation, we focused on the temperature and humidity correlation specifically which is represented by ρ_{hute} in the algorithm. For the inter-nodes correlation, we computed the correlation for the temperature in Sensors S1 and S2.

Algorithm 1 Pearson Data Correlation and Prediction Algorithm PDCP run on node S_0

Set $th_{cor_{inter}}$, $th_{cor_{intra}}$, th_{up}, th_{low}, n, $i = 1$, R_m, ρ_{te}, ρ_{hute}, RN
Compute $\rho_{te_{S_0S_i}}$ for each neighbor S_i
If $\rho_{te} > th_{cor_{inter}}$ AND $RN_{S_0} > RN_{S_i}$ **then** S_i Stops sending the temperature Values to the sink **end if**
Compute ρ_{hute}
If $\rho_{hute} > th_{cor_{intra}}$ **then** Compute th_{up}, th_{low} and R_m ; Compute the next humidity value **end if**

5 Experimental Results

In this section, we compare PDCP algorithm to a Bayesian approach in [ea17]. We used a MATLAB simulator with a meteorological dataset for the 8^{th}, 9^{th} and the 10^{th} of April 2020 from two sensor nodes deployed in Lille city, France (Lille airport S_1 and Lille city centre S_2) from Weather_Underground website which gathers data from a sensor network of different weather stations deployed around the globe *. For the inter-nodes correlation the temperature parameter was taken as the studied parameter between both sensors. For the intra-node correlation the temperature and humidity parameters in each sensor were selected. The sampling rate of the sensor node is set to 1 value each 30 minutes (by default). April 8^{th} is used for learning the values and thresholds. These parameters are used in the testing phase in the next two days. The temperature in those 3 days varied from 9 degrees Celsius as a minimum to 26 degrees Celsius as a maximum. The humidity varied from 30% to 100%.

5.1 Inter-Nodes data reduction

The threshold of correlation th_{cor} is set to 0.9, since in this part we need a very high correlation to stop sending one of the two values. In those 3 days, 144 values from each parameter are sensed at each sensor-node. Simulations show that the correlation coefficient for the whole first day is equal to 0.91, greater than the predefined threshold of correlation (0.9) for the inter-nodes correlation, and $R_m = 1.09$. This high correlation coefficient leads to send only one out of the two temperature values by S2. Fig 1a shows the difference between the real and the predicted values for the next 2 days (April 9 and 10). This difference does not surpass 1 degree Celsius.

5.2 Intra-Node data reduction

Different environmental values are sensed by each sensor node. We take the example of joint sensing of temperature and humidity on S1. We set $th_{cor} = 0.75$ since we compare two different parameters. On April

*. https ://www.wunderground.com

8, 2020, the correlation coefficient was equal to -0.8, which is greater than 0.75, so the humidity value is extracted from it. Fig 1b shows the estimated humidity for April 9 and 10, 2020. In those 2 days, 96 humidity values were captured, however, the node only sent 6 humidity values to the sink. The difference between the real and the predicted values did not surpass 7% which shows the reliability of our approach.

The numbers show that the intra-node part in PDCP algorithm when applied reduces the amount of sent data to the sink while maintaining the integrity and the accuracy of the data as shown in the figures above. A comparison with another method is drawn in the section below.

5.3 Intra-Node and Inter-Nodes Combination

In our approach, the sink is able to estimate temperature and humidity values of S1 from the first trend sent. Combining both approaches of PDCP algorithm (intra an inter node correlation) increases data reduction. Fig 1c shows the humidity prediction for S1, 7 values were sent. The maximum difference between the real and predicted humidity values never exceeds 7% of humidity. While applying the intra-node correlation on S2, the node sent only 7 humidity values to the sink in two days of predictions as shown in Table 1.

Table 1 draws the differences between our approach and a Bayesian inference approach from the literature [ea17] for the same scenario and parameters. As noticed from the numbers, they are neglecting any change in the humidity values which helps them to improve data reduction to 50% for the intra-node correlation but they lost some accuracy with a humidity standard deviation H_{SD} up to 10%. However, the inter-nodes correlation applied in PDCP gives us the edge to improve the percentage of data reduction to reach 70% with a better accuracy and a standard deviation H_{SD} of 7%.

TABLE 1: Amount of transmitted values per day by S1 and S2

Day	All data	PDCP	Bayesian [ea17]	Day	All data	PDCP	Bayesian
$S1_T$	96	0	96	$S2_T$	96	96	96
$S1_H$	96	7	0	$S2_H$	96	5	0
Total	384	108	192	H_{SD}	0%	7%	10%
Data Reduction	0%	70%	50%				

6 Conclusion and Future Work

In this paper, we proposed data reduction based on the Pearson correlation functions for WSN based agriculture monitoring. Our simulations show a reduction of more than 70% of the overall data which surpasses other approaches from the literature by more than 20%. As future work, this approach will be enhanced by being included in routing protocol for multi-hop scenarios. To further

Références

[ea15] K. Musaazi et al. Energy efficient data caching in wireless sensor networks : A case of precision agriculture. In *e-Infrastructure and e-Services for Developing Countries*, 2015.

[ea17] C. Razafimandimby et al. A Bayesian approach for an efficient data reduction in IoT. In *InterIoT*, 2017.

[ea18] G. Bou Tayeh et al. A distributed real-time data prediction and adaptive sensing approach for wireless sensor networks. *Pervasive and Mobile Computing*, 49 :62 – 75, 2018.

[RC19] G. Rajesh and A. Chaturvedi. Correlation analysis and statistical characterization of heterogeneous sensor data in environmental sensor networks. *Computer Networks*, 164, 2019.

[SM20] C. Salim and N. Mitton. K-predictions based data reduction approach in wsn for smart agriculture. *Computing*, pages 1–24, 2020.